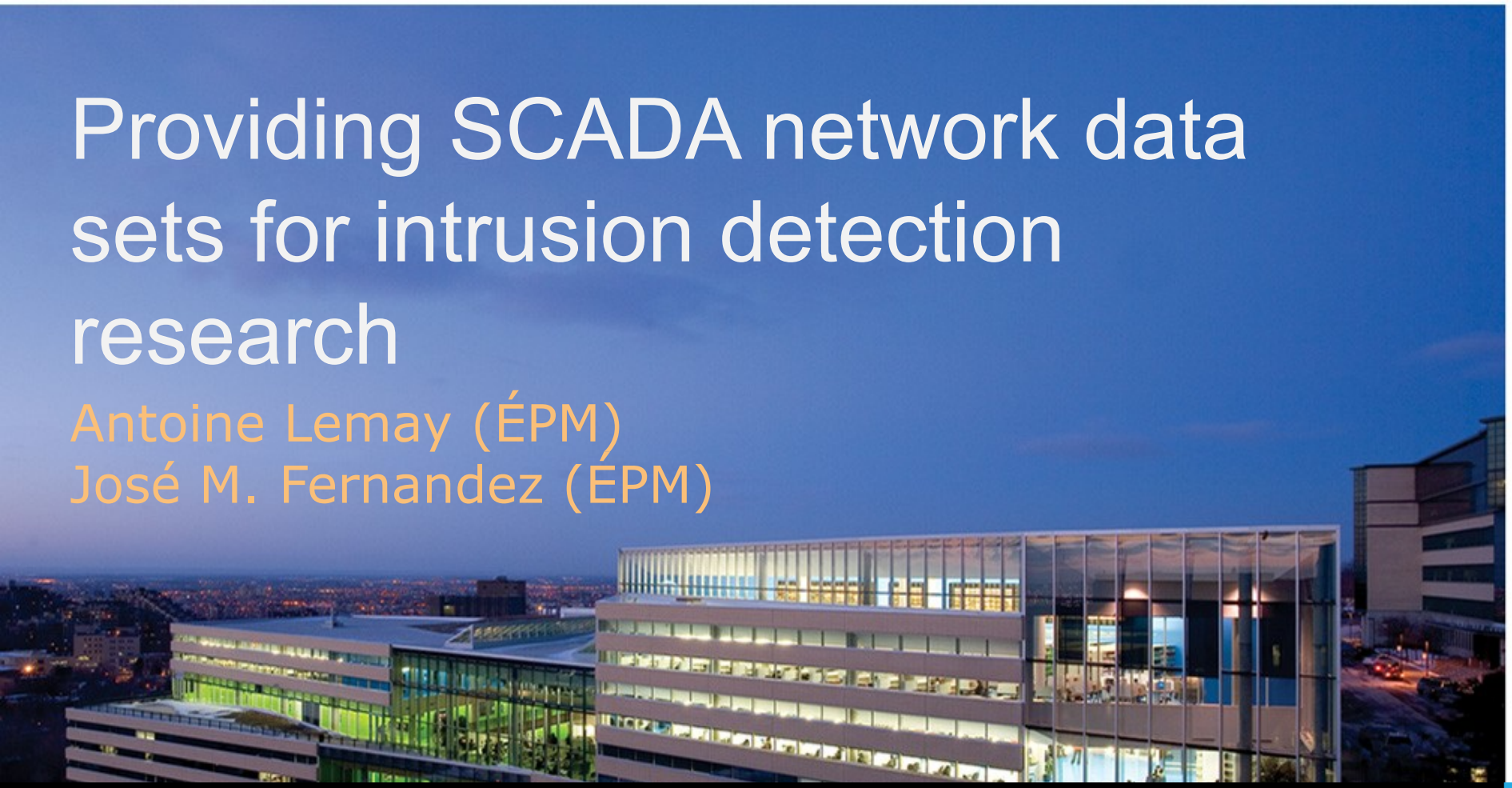
Providing SCADA network data sets for intrusion detection research

Antoine Lemay (ÉPM)
José M. Fernandez (ÉPM)



WORLD-CLASS ENGINEERING

POLYTECHNIQUE MONTRÉAL

PLAN

- Introduction to SCADA networks
- (Mis?)use of SCADA data sets in the literature
- Challenges
- Traffic generation infrastructure
- Data sets
- Conclusion and future work

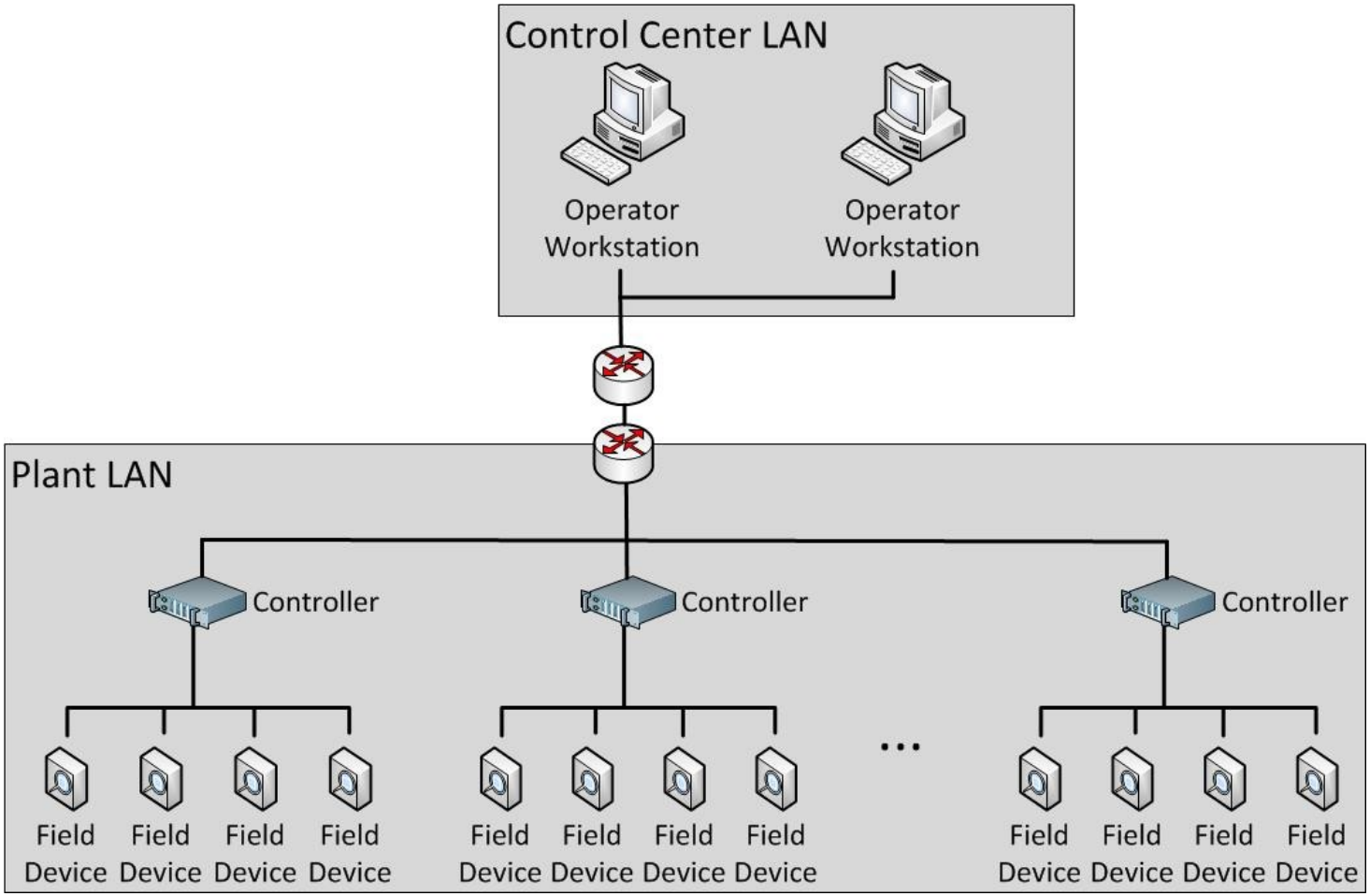
PLAN

- Introduction to SCADA networks
- (Mis?)use of SCADA data sets in the literature
- Challenges
- Traffic generation infrastructure
- Data sets
- Conclusion and future work

INTRODUCTION TO SCADA NETWORKS

- Supervisory Control And Data Acquisition (SCADA) is a form of industrial automation
  - Sensors collect data (DA)
  - Actuators allow remote control (C)
  - Centralized system
- Operators connect to the central data repository through Human Machine Interfaces (HMI)
  - Dedicated workstation for the operators
- Operator actions in the HMI are translated into SCADA network protocol packets
  - Changing the status of a valve in the HMI will trigger the sending of a Modbus write coil packet

TYPICAL SCADA ARCHITECTURE



PLAN

- Introduction to SCADA networks
- (Mis?)use of SCADA data sets in the literature
- Challenges
- Traffic generation infrastructure
- Data sets
- Conclusion and future work

SECSI

USE OF DATA SETS IN THE LITERATURE

- Hadžiosmanović et al. [4] research on anomaly detection in n-grams
  - Use of DigitalBond dataset
  - Emphasis on malformed packets
- Barbosa et al. [5, 6] research on intrusion detection using traffic from a water plant deployment
  - Use of real data
  - How do we establish the ground truth ?
- Valdes and Cheung [7] research on anomaly detection
  - Effects of using the wrong model for data (use of T-Test on non-normal data)

PLAN

- Introduction to SCADA networks
- (Mis?)use of SCADA data sets in the literature
- Challenges
- Traffic generation infrastructure
- Data sets
- Conclusion and future work

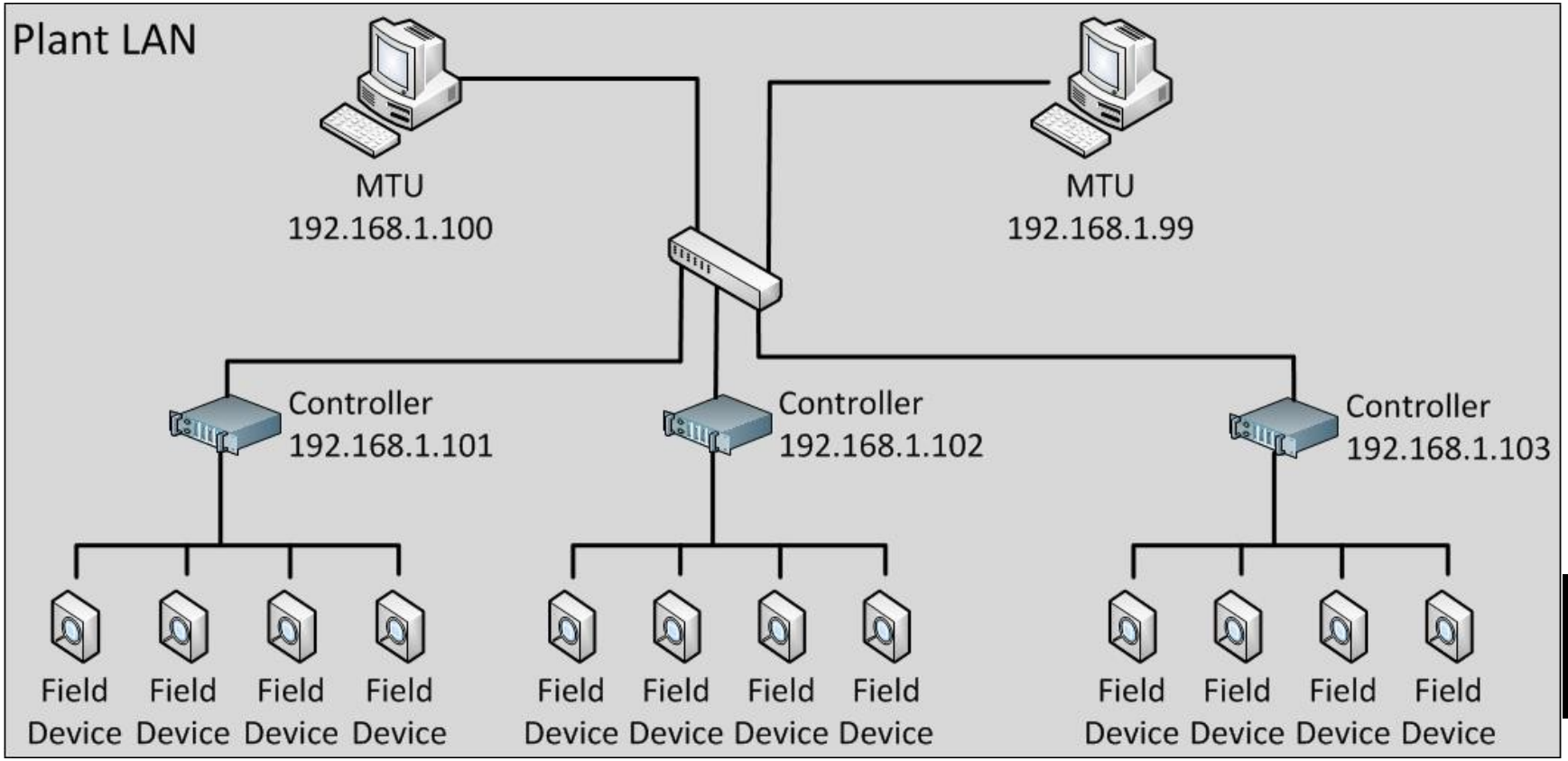
CHALLENGES

- Timing
  - Scaling attack traffic to polling timing
- Injecting noise in the traffic
  - Match information properties of the physical system
- Labeling
  - Definition of malicious (intent-based?)
  - Especially important for some of our attack types
  - Use of a connection-based label (all the packets of a connection that contains malicious traffic are labeled malicious)

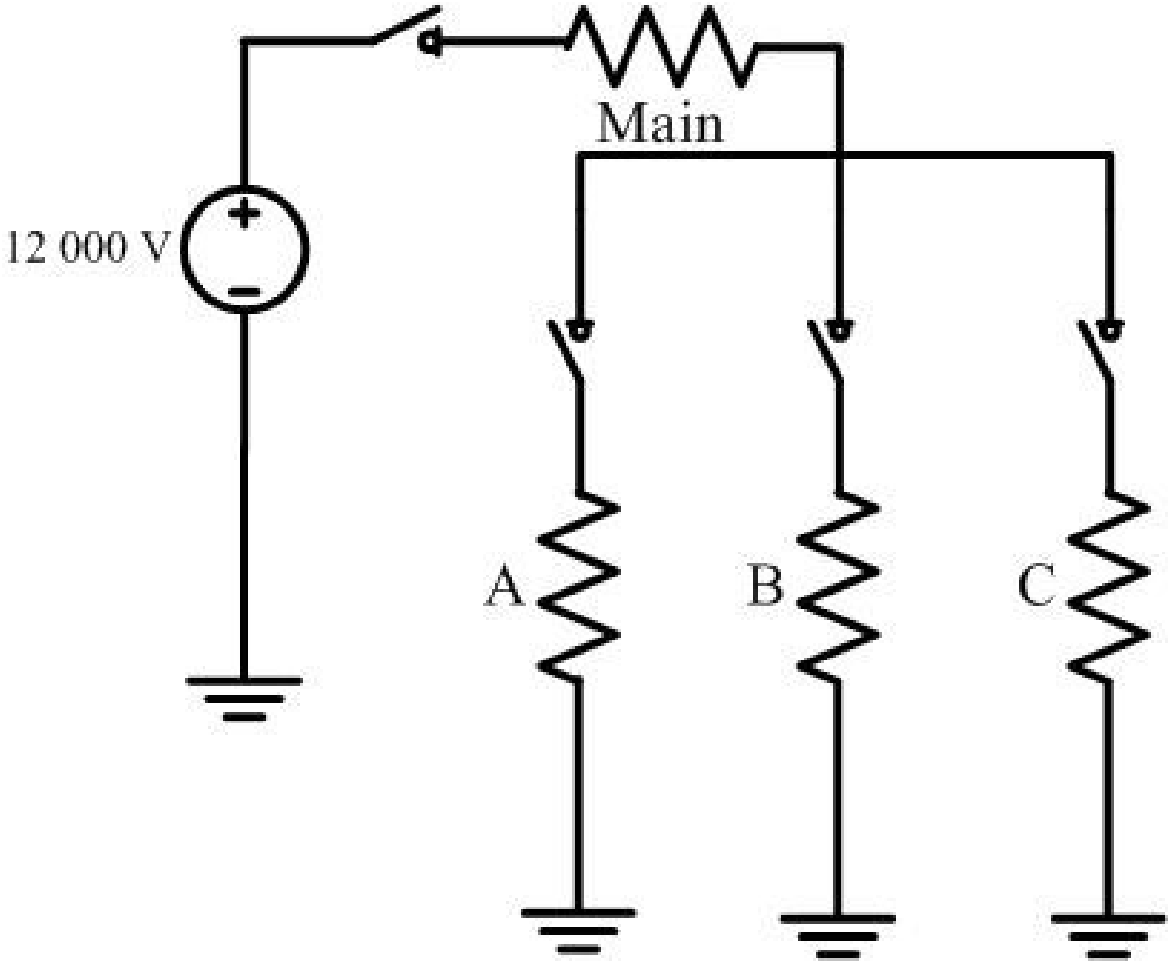
PLAN

- Introduction to SCADA networks
- (Mis?)use of SCADA data sets in the literature
- Challenges
- Traffic generation infrastructure
- Data sets
- Conclusion and future work

GENERATION OF THE TRAFFIC



GENERATION OF THE TRAFFIC



PLAN

- Introduction to SCADA networks
- (Mis?)use of SCADA data sets in the literature
- Challenges
- Traffic generation infrastructure
- Data sets
- Conclusion and future work

DATA SETS

- Two data sets
  - Modbus
  - Covert channel
- The Modbus data set uses variations of the default network
  - Attack traffic vs no attack traffic
  - Number of controllers
  - Polling interval
  - "Human" interaction or no interaction
- Covert channel data set includes only attack traffic using a different network
  - Based on Lemay, Fernandez and Knight [13]
- All data sets and associated labeling files available at :
- https://github.com/antoine-lemay/Modbus_dataset

DATA SET - MODBUS

Name	Description	Malicious activity?	Number of entries
Run8	1 hour of regular Modbus traffic including polling and manual operation - 2 MTU, 3 RTU and 10 seconds polling interval	No	72 186
Run11	1 hour of regular Modbus traffic including polling and manual operation - 2 MTU, 3 RTU and 10 seconds polling interval	No	72 498
Run1_6RTU	1 hour of regular Modbus traffic including polling and manual operation - 2 MTU, 6 RTU and 10 seconds polling interval	No	134 690
Run1_12RTU	1 hour of regular Modbus traffic including polling and manual operation - 2 MTU, 12 RTU and 10 seconds polling interval	No	238 360
Run1_3RTU_2s	1 hour of regular Modbus traffic including polling and manual operation - 2 MTU, 3 RTU and 2 seconds polling interval	No	305 932
Modbus_polling_only_6RTU	1 hour of regular Modbus traffic including polling only - 1 MTU, 3 RTU and 10 seconds polling interval	No	58 325

SECSI

DATA SET - MODBUS

Name	Description	Malicious activity?	Number of entries
Moving_two_files_Modbus_6RTU	3 minutes of regular Modbus traffic including polling only - 1 MTU, 6 RTU and 10 seconds polling interval	Yes	3 319
Send_a_fake_command_Modbus_6RTU_with_operate	11 minutes of regular Modbus traffic including polling and manual operation - 1 MTU, 6 RTU and 10 seconds polling interval. Also includes sending a Modbus write operation from a compromised RTU using Metasploit proxy functionality and the proxychains tool	Yes	11 166
Characterization_Modbus_6RTU_with_operate	5.5 minutes of regular Modbus traffic including polling and manual operation - 1 MTU, 6 RTU and 10 seconds polling interval Also includes sending a series of modbus read commands to characterize available registers from a compromised RTU	Yes	12 296
CnC_uploading_exe_modbus_6RTU_with_operate	1 minute of regular Modbus traffic including polling and manual operation - 1 MTU, 6 RTU and 10 seconds polling interval. Also includes sending an EXE file from a compromised RTU to another compromised RTU through a Metasploit meterpreter channel	Yes	1 426
6RTU_with_operate	5 minutes of regular Modbus traffic including polling and manual operation - 1 MTU, 6 RTU and 10 seconds polling interval. Also includes using an exploit (ms08_netapi) from a compromised RTU to compromise another RTU using Metasploit	Yes	1 856

DATA SET – COVERT CHANNEL

Name	Description	Malicious activity?	Number of entries
Channel_2d_3s	Modbus covert channel using the two least significant digits of three storage registers	Yes	383 312
Channel_3d_3s	Modbus covert channel using the three least significant digits of three storage registers	Yes	255 668
Channel_4d_1s	Modbus covert channel using the four least significant digits of one storage registers	Yes	414 412
Channel_4d_2s	Modbus covert channel using the four least significant digits of tw0 storage registers	Yes	266 387
Channel_4d_5s	Modbus covert channel using the four least significant digits of five storage registers	Yes	107 577
Channel_4d_9s	Modbus covert channel using the four least significant digits of nine storage registers	Yes	60 295
Channel_4d_12s	Modbus covert channel using the four least significant digits of twelve storage registers	Yes	44 977
Channel_5d_3s	Modbus covert channel using the five least significant digits of three storage registers	Yes	143 809

PLAN

- Introduction to SCADA networks
- (Mis?)use of SCADA data sets in the literature
- Challenges
- Traffic generation infrastructure
- Data sets
- Conclusion and future work

CONCLUSION

- Importance of the choice of data sets
  - Examples of errors induced by poor use of data sets in the literature
- Challenges associated with creating SCADA data sets
  - Matching polling intervals
  - Matching natural information properties of the traffic
  - Labelling malicious traffic
- Presented a data set to be used as a baseline for research
  - Freely available https://github.com/antoine-lemay/Modbus_dataset

FUTURE WORK

- Improving limitations
  - Diversity of information properties
  - Diversity of controller configuration
  - Diversity of protocols
- Path forward
  - Calibrating existing infrastructure with real network data
  - Adding more attacks
- Obstacles
  - Non-disclosure
  - Generation of “human” traffic

SECSI

WORKS CITED IN THIS PRESENTATION

[4] D. Hadžiosmanović, L. Simionato, D. Bolzoni, E. Zambon and S. Etalle, "N-Gram Against the Machine: On the Feasibility of the N-Gram Network Analysis for Binary Protocols," 15th International Symposium on Research in Attacks, Intrusion and Defenses (RAID) . pp. 354–373, Amsterdam, 2012.

[5] R. R. R. Barbosa, A. Pras and R. Sadre, "Flow whitelisting in SCADA networks," in *7th Annual IFIP Working Group 11.10 International Conference on Critical Infrastructure Protection*, Washington, 2013.

[6] R. R. R. Barbosa, R. Sadre and A. Pras, "A First Look into SCADA Network Traffic," in *Network Operations and Management Symposium* (NOMS)Maui, 2012.

[7] A. Valdes and S. Cheung, "Communication Pattern Anomaly Detection in Process Control Systems," in *IEEE Conference on Technologies for Homeland Security*, Waltham, MA, 2009.

[13] A. Lemay, J. M. Fernandez and S. Knight, "A Modbus command and control channel," in *IEEE Systems Conference* (SYSCON), Orlando, 2016.

SECSI

QUESTIONS